DOCUMENT RESUME

ED 362 534 TM 020 543

AUTHOR Clauser, Brian; And Others
TITLE The Effects of Score Group Width on the Mantel-Haenszel Procedure.
PUB DATE Apr 92
NOTE 24p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Ability; Comparative Analysis; Computer Simulation; *Item Bias; Reference Groups; *Scores; *Statistical Distributions
IDENTIFIERS *Mantel Haenszel Procedure; *Score Groups; Type I Errors

ABSTRACT

Previous research examining the effects of reducing the number of score groups used in the matching criterion of the Mantel-Haenszel procedure, when screening for differential item functioning, has produced ambiguous results. The goal of this study was to resolve the ambiguity by examining the problem with a simulated data set. The main results from this study call into question the preliminary recommendations of several other researchers, that four or more score groups are sufficient and produce stable results. Although considerable stability and very little Type I error was noted with equal ability distribution comparisons, with unequal ability distributions, the Type I error rate was substantially inflated. These results argue against the appropriateness of implementing the procedure by collapsing score groups. The current data suggest that more than modest reductions in the number of score groups cannot be recommended when the ability distributions of the reference and focal groups differ. One figure and five tables illustrate the discussion. (Contains 14 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 362 534

THE EFFECTS OF SCORE GROUP WIDTH ON THE MANTEL-HAENSZEL PROCEDURE

Brian Clauser
National Board of Medical Examiners

Kathleen M. Mazor and Ronald K. Hambleton
University of Massachusetts at Amherst

Abstract

Previous research examining the effects of reducing the number of score groups used in the matching criterion of the Mantel-Haenszel procedure, when screening for DIF, has produced ambiguous results. The goal of this study was to resolve the ambiguity by examining the problem with a simulated data set.

The main results from this study call into question the preliminary recommendations of several other researchers, that four or more score groups are sufficient and produce stable results. Although considerable stability and very little type I error was noted with equal ability distribution comparisons, with unequal ability distributions, the type I error rate was substantially inflated. These results argue against the appropriateness of implementing the procedure by collapsing score groups. The current data suggest that more than modest reductions in the number of score groups cannot be recommended when the ability distributions of the reference and focal groups differ.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

RONALD K. HAMBLETON

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM020543

THE EFFECTS OF SCORE GROUP WIDTH ON THE MANTEL-HAENSZEL PROCEDURE[1,2]

Brian Clauser
National Board of Medical Examiners

and

Kathleen M. Mazor and Ronald K. Hambleton
University of Massachusetts at Amherst

The identification of differentially functioning items remains a major concern for test developers. Although a variety of approaches have been documented (Hills, 1989; Scheuneman & Bleistein, 1989), no single approach has emerged as optimal. Considerable research is available comparing these approaches and examining their performance in various contexts. The present paper adds to this literature with an empirical assessment of the Mantel-Haenszel (MH) statistic.

The MH statistic tests the null hypothesis that the odds of a correct response to a given item is equal for members of the focal and reference groups after they have been matched on the ability of interest. This can be done using a valid external measure of the ability. However, such a measure is generally not available. More typically, this matching is carried out using the total test score as the criterion. The Mantel Haenszel α represents the sum of the odds ratios at each score level within the criterion, weighted by the number of examinees at that score level. This allows for k+1 score groups, where k is the number of items on the criterion test.

In their original paper recommending the MH statistic as a DIF detection procedure, Holland and Thayer (1988) assume that k+ 1 score groups (i.e., the maximum possible number of score groups given the data) will be used in the

---

[1]Laboratory of Psychometric and Evaluative Research, Report No. 226. Amherst, MA: University of Massachusetts, School of Education.

[2]Paper presented at the meeting of AERA, San Francisco, 1992.

matching criterion. There is obvious appeal in the notion that matching should be carried out as finely as the data allow. Any alternative would seem to allow group differences within matching levels to unnecessarily diminish the control such conditioning is intended to produce (Angoff, 1993). More recently, Donoghue, Holland, and Thayer (1993) reported results based on a simulation study suggesting that, even when using the maximum available number of score categories, problems may occur when matching is based on very short tests. They reported that matching based on tests of four or nine items produced unsatisfactory results. When tests of 19 or 39 items were used, the statistic performed acceptably.

Nonetheless, for a number of reasons, researchers and practitioners have been interested in the possible advantages of using this procedure with fewer than the maximum number of score groups in the matching criterion. Raju, Bode, and Larsen (1989) suggest that if the power of Scheuneman-type Chi-square tests (Scheuneman, 1979) increases as the number of score groups decreases, then examination of the MH statistic under these conditions would seem important. Additionally, Hills (1989) highlights one of the advantages of the MH procedure as being its usefulness with relatively small examinee samples. However, score levels that appear in only one group (i.e., focal or reference) are dropped from the calculations. Such a loss is most critical with small samples. This problem can be reduced or eliminated if score groups are combined.

Both Raju, et al. (1989) and Wright (1986) have provided data on the effects of varying the number of score groups used. Unfortunately for the practitioner, their results may raise as many questions as they answer. Raju and his associates conclude that "4 or more score groups yield stable α estimates with the MH technique" (p. 11). However, these conclusions are

based on comparisons of 2, 4, 6, 8, and 10 score groups. These alternatives are compared to each other, but not to the total possible number which, for the 40-item test they examined, would have been 41. Interpretation of the results is further complicated by the surprisingly high numbers of items identified as differentially functioning. In the Black versus White comparison (at the .05 level), 16 out of a possible 40 items were identified with 10 score groups and 24 were identified with two score groups. The authors attribute this "inflated type I error" to the "large number of Chi-square values involved" in the analysis (p. 12), and suggest a procedure such as the Bonferroni method to control it. Given that approximately half of the items on the test had been identified as differentially functioning, it would seem reasonable to consider other explanations. But because the data are actual test results, it is impossible to know which items are correctly identified and which represent type I error.

Wright's (1986) work in this area is interesting on two counts. First, he adds an important dimension to the study by comparing different numbers of score groups under conditions of different sample sizes. Secondly, he presents results which seem to conflict with the results of Raju et al. Wright suggests that six score groups are inadequate when compared to the data produced with 61 score groups. As with the Raju et al. paper, these results are somewhat ambiguous because the analysis was conducted on actual test results, allowing no clear means to differentiate between increased power and increased type I error.

The present research attempts to eliminate some of the ambiguity found in the results described above by examining the score group variable with simulated data. This research follows Wright's lead by varying sample size as well.

Method

The current study uses simulated data produced using DATAGEN (Hambleton & Rovinelli, 1973), a computer program to simulate examinee item response data fitting a one-dimensional logistic model. To produce the test into which the simulated DIF items were placed, a- and b-parameters were taken from 70 items from the 1985 administration of the Graduate Management Admission Test (Kingston, Leary, & Wightman, 1988). These values are shown in Table 1. These parameters were chosen to more closely approximate conditions found in practice. Ten studied items were then added to the 70 to make a total test of 80 items. This test length was chosen because it is within the range occurring in typical standardized testing situations (e.g., achievement subtests) and yet long enough to reduce the instability which can be associated with Mantel-Haenszel results for shorter tests. It also allowed for a substantial number of items for study without making the percentage of DIF items in the test greater than that which has been routinely identified in actual tests.

- -

Insert Table 1 about here

- -

The c-parameters for all items were set at 0.20. The a-parameters for the studied items were set at 0.25, 0.60, 0.90, or 1.25. This approximated the range of values found in the estimated GMAT parameters. These values were crossed with five levels of simulated DIF represented by differences in the b-parameters for focal and reference groups of 0.00, 0.25, 0.50, 1.00, and 1.50. These 20 item parameter combinations (four levels of item discrimination x five levels of b-value difference) were then crossed with five levels of item

difficulty (with reference group b-parameter values of -2.50, -1.00, 0.00, 1.00, and 2.50). To allow for this number of studied items (100), ten simulated tests were used, with each containing 10 studied items. (Note that 20 of these studied items did not display DIF. Those with no difference in the reference and focal group b-parameters were included to allow for examination of the type I error rate associated with the studied conditions.)

The examinee item responses for the 70 core items were held constant for the ten test simulations to prevent chance differences in these responses from influencing the effects under study. Responses were produced for 2,000 examinees in each group. Because individual examinee response patterns produced by DATAGEN are random, smaller examinee samples were produced by selecting the first 1,000, 500, 200, or 100 examinees from each group.

Ability distributions were created for the reference and focal groups so they would be equal and normally distributed with a mean of 0.0 and a standard deviation of 1.0. This arrangement is similar to that which is often encountered when using the MH statistic to assess for male-female differences. The simulations were then repeated using distributions that differed by 1.0 standard deviation. The distribution for the reference group remained as described above. The distribution for the focal group had the same shape with a lower mean. This arrangement was intended to simulate conditions found in other types of reference-focal group comparisons (see, for example, Hambleton & Rogers, 1989; Raju, Bode, & Larsen, 1989).

The MH statistic was then calculated for each item of the above data sets. The form of the statistic used was the two-step procedure recommended by Holland and Thayer (1988). With this procedure, items identified as displaying DIF (at a .01 level of significance) on a first MH run are removed from the matching criterion used for the second MH run. At each of five

sample sizes (i.e., 2,000, 1,000, 500, 200, 100), the calculations were replicated with five different numbers of score groups (i.e., 81, 20, 10, 5, and 2 score groups). Score groups were created to be as close as possible to equal in width. This equality was based on the number of possible scores used in the matching criterion. When five score groups were used, the intervals would be 0-15, 16-31, 32-47, 48-63, and 64-80.

- -

Insert Tables 2 and 3 about here.

- -

Results

Table 2 summarizes the Mantel-Haenszel results across sample sizes and number of score groups for the examinee samples with equal ability distributions. Each entry represents the total percent of items that were identified out of a possible 80 (uniform) DIF test items. This was accomplished by adding the DIF items identified in each of the ten tests. As the results indicate, statistical power increases as the sample size increases. Also, the results reflect substantial stability across the number of score groups used in the MH calculations. Table 3 replicates these results for the unequal ability distributions. By contrast, under this condition, decreasing the number of score groups was associated with substantial increases in the percents of DIF items identified. Across the five sample sizes, the average percent of DIF items detected increased by 16.4% as the number of score groups was decreased from 81 to 2.[1]

Table 4 provides the results across the five sample sizes for the data included to allow examination of type I errors. With equal ability distributions, and regardless of sample size or number of score groups, the type I error rate was low. A total of only six items (across 25 analyses)

were incorrectly identified. This closely approximates the nominal type I error rate (Nominal = .01, actual = .012). With unequal ability distributions, the results were very different. The type I error rate reached 75% with a large sample and a very small number (2) of score groups. In contrast, the type I error rate was very low, regardless of sample size, with 20 or 81 score groups.

Table 5 shows the type I error rates associated with the 70 core items. The values reflect the mean across each of the ten runs per condition (i.e., one run for each of the ten sets of studied items). These results vary somewhat from those presented in Table 4, but are in general quite similar. Those in Table 4 may provide a more appropriate base rate for comparison with the simulated DIF items while the results in Table 5 provide support for the notion that these results can be generalized to actual data.

- -

Insert Tables 4 and 5 about here

- -

Discussion

The data suggest very clearly that the extent to which there is anything to be lost or gained by varying the number of MH score groups depends on the characteristics of the examinee sample under comparison. For relatively large sample sizes with very similar ability distributions, there is considerable stability across the number of score groups. With 2,000 examinees in each group, no change in the items identified was noted between 81 and 2 score groups. With 1,000 examinees in each group, only two additional items were identified while reducing the score groups from 81 to 2. For smaller samples (still with equal ability distributions), the number of additional items

identified remains low. However, the percentage increase these gains represent is more impressive. With a sample size of 100, the change from 81 to 2 score groups translated into a 27% increase in the number of items correctly identified while increasing from 11 to 14 items.

Although the gains in power associated with reducing the number of score groups appear to be modest, they do not seem to be associated with an inflation in the type I error rate. In general, these results seem to suggest that when examinee groups are well matched in terms of ability distributions, there may be an advantage to reducing the number of score groups used, particularly if the comparison is based on very small samples. Unfortunately, the utility of this finding for measurement practice is limited in two ways. First, such well-matched ability distributions tend to be typical of comparisons between groups such as males and females. It is generally not difficult to collect larger samples from among these groups, even during piloting of a test. Second, the sample sizes for which a reduction in number of score groups would produce a substantial benefit are of such low power (below 20% in this study) that they are below the minimum recommended (Mazor, Clauser, & Hambleton, 1992) and should be avoided whenever possible.

When samples with unequal ability distributions are considered, the advantages associated with using fewer than the maximum possible number of score groups are more apparent. With examinee groups of 2,000, a 7% increase (i.e., 68% to 73%) when moving from 81 to 5 score groups is observed. Such increases are consistent across all but the smallest sample sizes. Unfortunately, the usefulness of this apparent advantage is even more limited than with equal ability distribution comparisons. As the results in Table 4 indicate, reduction in the number of score groups is associated with a substantial increase in type I error.

This type I error rate calls into question whether using the MH procedure with fewer than the maximum number of score groups produces increased statistical power or results in a random identification of additional items without distinguishing between items that function differentially and those that do not. A direct comparison of type I error rate and the rate of identification for previously unidentified DIF items is not encouraging. For a sample size of 1,000, the identification rate for DIF items associated with moving from 81 to 5 score groups increased about 21%. The increase in moving from 81 to 2 score groups was 36%. This compares to a type I error rate moving from 81 to 5 score groups of 20%, and 50% when moving from 81 to 2 score groups. Comparisons for other samples show similar patterns. Although these results clearly do not provide a definitive answer, they suggest that considerable caution should be used in interpreting the results of this application of the MH procedure. They suggest that the type I error rate is greatest under the same conditions that the MH identification rate is highest. High type I error rates were also noted among the 70 core items making up the tests into which the studied items were placed. Although the numbers are not as extreme, the patterns are similar.

The results do provide clues as to an explanation of the type I error rate. The fact that these errors are inflated only with unequal ability distributions suggests that combining score groups under such conditions may result in contamination of the matching criterion. The procedure assumes that all examinees within a given score group are of equal ability. As the score group width is extended, with unequal ability distributions, this assumption will not be met. The result is an invalid matching criterion allowing impact to be misinterpreted as DIF. Use of such a criterion could lead to falsely identifying acceptable items. Figure 1 provides a graphic representation of

the extent to which reducing the number of score groups in the matching criterion results in inaccurate matching. Obviously, when the maximum possible number of score categories are used, the mean score within each category for focal and reference members is equal. As the number of categories is reduced to ten, this equality begins to break down, but only to a small extent and only at the highest and lowest score categories. When it is further reduced to two categories, focal and reference groups may have substantially different means in both categories. This would adversely affect the validity of the matching for all examinees.

The results of an examination of the parameters of those items which were correctly identified as displaying DIF were, in general, consistent with previous findings reported by Clauser, Mazor, and Hambleton (1991). Items with moderate to high a-parameters, items with greater differences in b-parameters between groups, and items with medium to low reference group b-parameter values were most likely to be identified. The pattern for studied items displaying type I error was similar. For the unequal ability distribution comparison, only 4% of the instances of the type I error were in items with a-parameters of 0.25. Similarly, only 4% were in items with b-parameter values of 2.50. The other three a-parameter values (1.25, .90, and .60) accounted for 28%, 36%, and 30%, respectively. The other four b-parameter values (-2.50, -1.00, 0.00, 1.00) were associated with 18%, 28%, 18%, and 32%, respectively.

In generalizing the results of this study to actual data sets, the reader should note that, although various sample sizes were examined, the focal and reference group samples were always equal. This equality is typical in practice when making male-female comparisons but may not be the case when comparisons are made of various ethnic groups. Previous research examining

the sample size variable (Clauser, 1993) suggests that, although the power of the statistic may change when reference and focal groups are unequal, the pattern of results remains unchanged. Nonetheless, some caution is appropriate in making generalizations to such conditions.

It should also be noted that only one approach to collapsing score groups was examined. In this research, collapsing was carried out to produce score groups of equal width. Alternatively, the divisions could have been made so as to place equal numbers of examinees in each group or could have been limited to the extreme ends of the ability scale, with the single purpose of avoiding the situation in which score categories existed which lacked representation from both focal and reference groups. Again, caution must be taken in generalizing the results of this study to those conditions.

For the practitioner, the results of this study suggest that more than a modest reduction in the number of score groups used cannot generally be recommended. The use of as few as four groups (as has been recommended with Scheuneman-type Chi-square tests) is not justified with the Mantel-Haenszel procedure. Increased sample sizes are a clearly preferable means of increasing the power of the statistic. In cases where this is impossible, the technique of decreasing the number of score groups may be helpful. It should, however, be used with considerable caution because of the substantial type I error rate that may result.

End Notes

1. It should be noted that nominal and actual sample sizes and number of score groups may vary under some conditions. As described above, score levels that appear in only one group (focal or reference) are dropped from the calculations. The computer program used to calculate the MH statistic provides cell counts for all items identified as having a significant MH value. Examining a sample of these items across conditions indicated that, in general, the discrepancy between the nominal and actual sample sizes was small. With two score groups there

was no difference. With five or 10 score groups, the difference was generally under 1% of the total sample. This was true for both equal and unequal ability distributions. When the 20 or 81 score groups were used, the discrepancy remained under 3% for the equal ability distribution conditions involving samples of 1,000 per group or more. With a sample of only 100 per group, it remained under 3% with 20 score groups but increased to 10% with 81 score groups. With an unequal ability distribution for focal and reference groups, less than 2% of the examinee sample was excluded with 20 and less than 3% with 81 score groups, with samples of 1000 examinees per group or more. With the smallest samples examined (i.e., 100 per group), these rates increased to less than 10% for 20 score groups and less than 16% for 81 score groups. The numbers of excluded examinees tended to be relatively evenly split between focal and reference groups.

In addition to the nominal and actual sample sizes varying, the actual number of score groups used varied from the nominal number under some conditions. Because Holland and Thayer's (1988) two-step MH procedure was used, when items were identified on the first run as displaying DIF, they were removed from the matching criterion for the second run. This had no impact on the number of groups used when the nominal number was 2 through 20. When the nominal number of score groups was 81, 81 score groups were used for the first run, but as few as 73 were used for some of the second runs.

References

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), Differential item functioning: Theory and practice. Hillsdale, NJ: Lawrence Erlbaum Associates.

Clauser, B., Mazor, K., & Hambleton, R. K. (1991, April). An examination of item characteristics on Mantel-Haenszel detection rates. Paper presented at the meeting of the National Council on Measurement in Education, Chicago.

Clauser, B. E. (1993). An examination of factors influencing the performance of the Mantel-Haenszel procedure in identification of differential item functioning. Unpublished doctoral dissertation, University of Massachusetts at Amherst.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P. W. Holland & H. Wainer (Eds.), Differential item functioning: Theory and practice. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. Applied Measurement in Education, 2, 313-334.

Hambleton, R. K., & Rovinelli, R. J. (1973). A Fortran IV program for generating examinee response data from logistic test models. Behavior Science, 17, 73-74.

Hills, J. R. (1989). Screening for potentially biased items in testing programs. Educational Measurement: Issues and Practice, 8, 5-11.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), Test validity. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kingston, N., Leary, L., & Wightman, L. (1988). An exploratory study of the applicability of item response theory methods to the Graduate Management Admissions Test (GMAC Occasional Papers). Princeton, NJ: Graduate Management Admissions Council.

Mazor, K., Clauser, B., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. Educational and Psychological Measurement, 52, 443-451.

Raju, N.S., Bode, R. K., & Larsen, V. S. (1989). An empirical analysis of the Mantel-Haenszel statistic for studying differential item performance. Applied Measurement in Education, 2, 1-13.

Scheuneman, J. D. (1979). A method for assessing bias in test items. Journal of Educational Measurement, 16, 143-152.

Scheuneman, J. D., & Bleistein, C. A. (1989). A consumer's guide to statistics for identifying differential item functioning. Applied Measurement in Education, 2, 255-275.

Wright, D. J. (1986, April). An empirical comparison of the Mantel-Haenszel and standardization methods of detecting differential item performance. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco.

TABLE 1

IRT Parameters for 70 Simulated Items

Item	Parameter b	Parameter a	Item	Parameter b	Parameter a	Item	Parameter b	Parameter a	Item	Parameter b	Parameter a
1	0.44	0.64	21	-2.25	0.51	41	-2.09	0.34	61	1.24	0.61
2	-0.42	0.75	22	-1.30	0.61	42	0.55	1.27	62	1.41	0.75
3	1.39	1.04	23	1.23	0.82	43	-0.19	0.33	63	0.09	0.85
4	-1.17	0.47	24	-2.08	0.46	44	1.74	1.21	64	0.59	0.96
5	0.28	0.61	25	0.88	0.47	45	0.23	0.40	65	0.75	0.59
6	-1.47	0.32	26	0.06	0.44	46	-1.05	0.58	66	-0.78	0.63
7	0.37	0.72	27	-2.41	0.27	47	0.73	1.30	67	0.85	1.16
8	0.97	0.76	28	3.47	0.67	48	0.19	0.33	68	-1.70	0.43
9	-1.11	0.15	29	1.43	1.10	49	1.15	0.39	69	-0.23	0.43
10	-1.13	0.28	30	-1.23	0.59	50	0.51	0.78	70	0.24	0.79
11	0.22	1.00	31	1.40	0.95	51	0.51	0.55			
12	-1.07	0.30	32	2.27	0.70	52	0.80	0.53			
13	0.03	0.93	33	0.26	0.71	53	0.16	0.91			
14	-0.18	0.83	34	2.26	1.30	54	-0.12	0.84			
15	-1.61	0.54	35	0.99	0.22	55	1.29	0.10			
16	-0.91	0.40	36	-1.73	0.37	56	0.07	0.69			
17	0.12	0.34	37	0.64	0.49	57	-0.47	0.44			
18	-1.20	0.38	38	-1.12	0.57	58	-0.45	0.53			
19	0.94	0.73	39	-0.91	0.43	59	1.61	0.69			
20	1.22	0.42	40	0.33	1.05	60	-2.77	0.24			

TABLE 2

Percent of Items Identified by the Mantel-Haenszel Statistic* with Equal Ability Distribution Groups (Out of 80 Items)

Sample Size/ Group	Number of Score Groups 2	5	10	20	81
2,000	73%	73%	71%	73%	73%
1,000	65	65	64	64	63
500	51	54	49	49	51
200	29	26	24	24	26
100	18	16	16	16	14

*$p < .01$

TABLE 3

Percent of Items Identified by the Mantel-Haenszel Statistic* With Unequal Ability Distribution Groups (Out of 80 Items)

Sample Size/ Group	Number of Score Groups 2	5	10	20	81
2,000	85%	73%	69%	69%	68%
1,000	79	70	66	63	58
500	66	53	45	45	41
200	38	29	25	25	24
100	21	16	14	14	16

*p<.01

TABLE 4

Percent of Studied (Non-DIF) Items Identified by the Mantel-Haenszel Statistic* (Out of 20 Items)

Ability Distributions	Sample Size/ Group	Number of Score Groups 2	5	10	20	81
Equal	2,000	0%	0%	5%	0%	0%
	1,000	0	0	0	0	0
	500	0	0	0	0	0
	200	0	5	5	5	5
	100	5	0	0	0	0
Unequal	2,000	75%	30%	10%	5%	5%
	1,000	50	20	0	0	0
	500	20	5	0	0	0
	200	10	0	0	0	0
	100	5	0	0	0	0

*$p < .01$

TABLE 5

Percent of Type I Error for 70 Core Items
Averaged Across Ten Runs

Ability Distributions	Sample Size/ Group	Number of Score Groups				
		2	5	10	20	81
Equal	2,000	1.4%	1.0%	0.7%	0.9%	0.6%
	1,000	0.0	0.0	0.0	0.0	0.0
	500	0.0	0.0	0.0	0.0	0.0
	200	0.6	0.0	0.0	0.0	0.1
	100	0.0	0.0	0.0	0.0	0.0
Unequal	2,000	75.4%	20.4%	3.7%	1.6%	1.4%
	1,000	45.9	9.6	1.9	1.4	1.4
	500	11.9	4.6	1.4	1.4	1.4
	200	5.3	0.4	0.0	0.0	0.0
	100	0.1	0.0	0.0	0.0	0.0

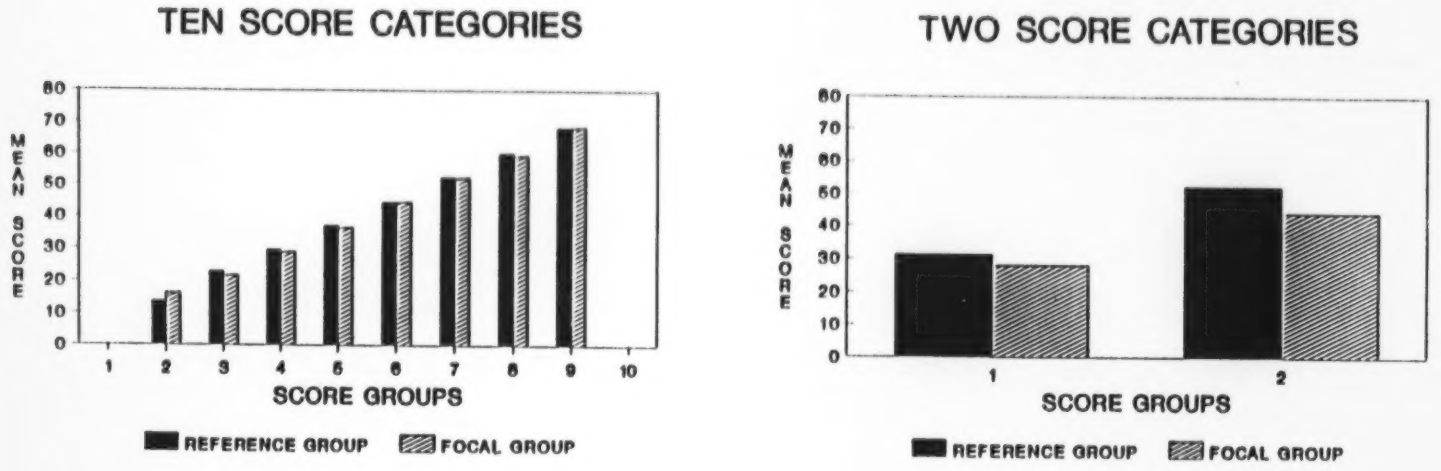

Figure 1. These graphs represent the mean score within each score category for reference and focal groups simulated to have unequal ability, with ten and two score categories.